Санкт-Петербургский государственный университет

Ревегук Иван Вячеславович

Приложение методов глубокого обучения к анализу белковых молекул

Выпускная квалификационная работа
по направлению подготовки “биология”

Работа выполнена в лаборатории
Функциональной Геномики
Московского Физико-Технического Института (МФТИ)
(зав. лаб. к.б.н. Скоблов М.Ю.)

Научный руководитель:
Стефанов В. Е.

Научный консультант
Корвиго И. О.

Санкт-Петербург 2018

Содержание

1 Введение

Из всех типов пост-трансляционных модификаций (ПТМ) изучению фосфорилирования было отведено наибольшее внимание. Фосфорилирование играет ключевую роль в качестве "молекулярного переключателя" во множестве типов регуляторных процессов в эукариотической клетке [1–3].

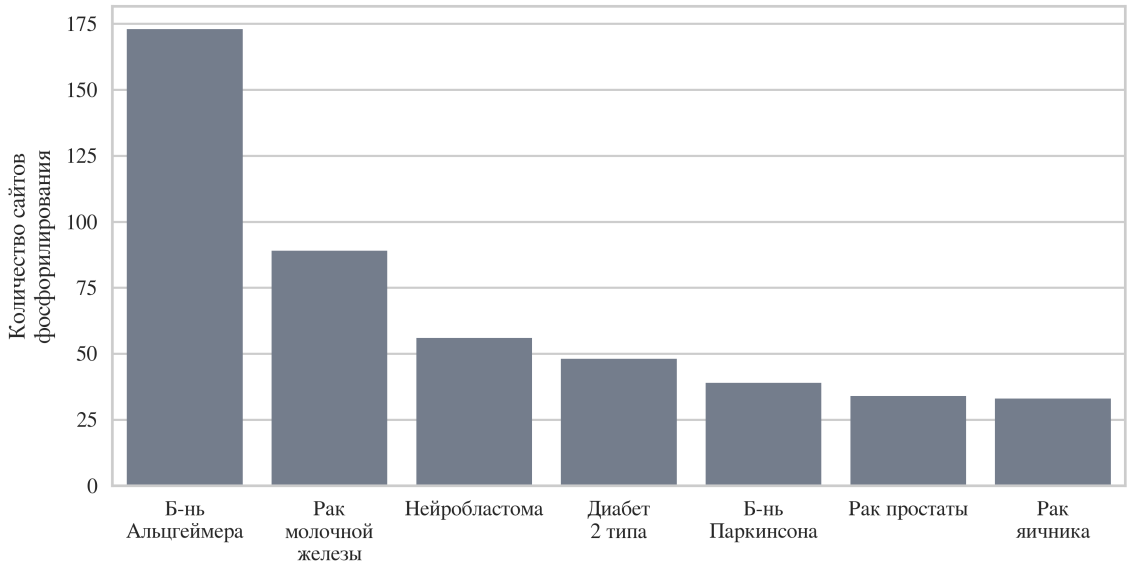
Для экспериментального изучения фосфорилирования и других ПТМ за последние три десятилетия учеными было разработано множество методов. Их можно разделить на две группы: низкопропускные и высокопропускные [4]. Объектом исследования низкопропускных методов является, как правило, один или несколько интересующих исследователя белков, в то время как высокопропускные методы, такие как тандемная масс-спектрометрия, позволяют проводить изучение ПТМ, а в частности, фосфорилирования полноценных клеточных протеомов.

В результате широкомасштабного использования высокопропускных экспериментальных методик за последние годы было накоплено большое количество фосфопротеомных данных. Последние сегодня доступны для широкому кругу пользователей посредством открытых баз данных в сети Интернет. Несмотря на то, что большинство данных были вручную извлечены экспертами из соответствующих научных статей, лишь немногие из баз данных в достаточной мере обеспечивают свое содержимое сопроводительной информацией, касающейся экспериментального базиса, использованного для обнаружения того или иного сайта ПТМ. Частое отсутствие подобных аннотаций в совокупности с аккумуляцей экспериментальных результатов, значительно отягощает оценку качества записей. Особенно явно это проявляется в случае использования вычислительных методов идентификации сайтов ПТМ, так как подобные методы базируются в подавляющем большинстве случаев на алгоритмах машинного обучения, эффективность которых напрямую зависит от качества выборки, используемой для обучения модели.

Точное предсказание сайтов ПТМ при помощи алгоритмов машинного обучения является крайне желаемой альтернативой экспериментальным техникам по причине дешевизны, доступности, легкости в использовании. На протяжении последних двадцати лет множество усилий было затрачено на создание подобных моделей, и сегодня на сайте `Omictools.org` [5] предлагается более пятидесяти моделей, подходящих для задачи предсказания сайтов фосфорилирования. Тем не менее, в данном списке отсутствует предиктор, способный к киназо- и таксоно- неспецифичным предсказаниям. Помимо этого, зачастую авторы прекращают поддержку моделей по истечении нескольких лет после публикации, что делает работу с ними попросту невозможной. В силу того, что машстаб протеомных исследований со временем увеличивается, необходимой является и способность предиктора работать с большими массивами данных.

Поэтому создание киназо-неспецифичного предиктора, работа которого основана на

Рис. 1: Количество сайтов фосфорилирования, мутации в которых соответствуют заболеваниям человека.



Использованы данные с PhosphoSitePlus. Отображены только заболевания, для которых описано более 30 сайтов.

современных алгоритмах машинного обучения, способного к максимально точным предсказаниям, и не ориентированного на работу в рамках узкой филогенетической группы является востребованной на сегодняшний день задачей. Подобная модель могла бы быть полезна в широком кругу актуальных в современной науке проблем: от дизайна лекарств и тестирования гипотез, основанного на результатах предсказания, до различного рода широкомасштабных статистических и филогенетических исследований. Кроме того, множество заболеваний человека ассоциировано с мутациями в области сайтов фосфорилирования. Так, в созданной экспертами ресурса PhosphoSitePlus на основе научных статей регулярно обновляемой выборке на момент написания данной работы указано более 1100 мутаций в области сайтов фосфорилирования, приводящих к 208 различным заболеваниям человека (см. Рис. 1). Создав модель, важно также проанализировать, насколько эффективность её работы зависит от критериев отбора, предъявленных к записям из открытых баз данных в процессе формирования тренировочных выборок. Также, с целью минимизации потенциальных ложных результатов предсказаний, в силу отсутствия таксономической специфичности, важно отдельно указать, для каких таксономических групп созданная модель работает наиболее эффективно.

2 Обзор литературы

На данный момент времени существует множество вычислительных моделей, осуществляющих предсказания тех или иных сайтов ПТМ [6]. Так как фосфорилирование является наиболее изученной ПТМ, ему посвящено большинство экспериментальных данных. Как следствие, параллельно с накоплением последних происходила разработка вычислительных методов для предсказания сайтов фосфорилирования.

Среди существующего множества предикторов, лишь четыре были найдены в той или иной степени соответствующими заявленым во Введении критериям, а также активно поддерживаемыми на данный момент, и способными к работе с относительно большими объемами данных. А именно GPS3.0 [7], MUSITE [8], PHOSFER [9] и PhosPredRF [10].

Таблица 1: Модели, использовавшиеся для сравнения и основная информация о них.

Название, год публикации	Киназо-неспецифичность	Размер тренировочной выборки	Базы данных	Модели для сравнения
GPS, 2008	Нет	$\sim 3 * 10^3$	P.ELM6.0	Scansite; KinasePhos1.0; KinasePhos2.0; NetPhosK; pkaPS
Musite, 2010	Да*	-	P.ELM8.2; Swiss-Prot; PhosphoPep; PhosPhAt; TAIR	-
PHOSFER, 2013	Да	$\sim 62 * 10^3$	Uniprot; TAIR; Phospho.ELM; PhosphoSitePlus; P^3DB	PhosPhAt; PlantPhos
PhosPredRF, 2015	Да	$\sim 6 * 10^3$	P.ELM9.0; PPA 3.0	GPS 2.1; NetPhos; PPRED; Musite; PhosphoSVM

* - Musite способен как к киназо-специфичным, так и к киназо-неспецифичным предсказаниям.

GPS (Group-based Prediction System) среди данного списка является единственным в чистом виде киназо-специфичным предиктором; также им можно пользоваться в виде отдельного приложения с графическим интерфейсом. GPS основан на разработанным

автором алгоритме - Group-based Phosphorylation Scoring Method Algorithm. Данный алгоритм использует широко известные в биоинформатике матрицы, описывающие частоту замен одного элемента последовательности на другой - BLOSUM62. При помощи последней происходит поиск в пределах последовательности сайтов подпоследовательностей длиной в 7 аминокислотных остатков и наиболее соответствующих им сайтов фосфорилирования. Другими словами, в каждую 7-подпоследовательность вносятся случайные мутации, и после внесения каждой мутации в попарной манере на основе BLOSUM62 вычисляется коэффициент сходства между ней и известными сайтами фосфорилирования. В GPS2.0 для сравнения используются сайты фосфорилирования 71 группы протеин-киназ. Алгоритм был разработан авторами в конце прошлого века, и это единственный найденный предиктор, никак не использующий методы машинного обучения. Публикация, сопутствующая выходу GPS версии 3.0, отсутствовала на момент написания, поэтому информация об используемых для тренировки и оценки модели выборках была взята из публикации, соответствующей версии 2.0. GPS2.0 был натренирован и протестирован на приблизительно трех тысячах сайтах фосфорилирования, в основном принадлежащих белкам млекопитающих, взятых из базы данных P.ELM (6.0).

Musite, как и GPS3.0 доступен в виде отдельного приложения. Программа может быть использована как для киназо-специфичных предсказаний, так и для киназо- неспецифичных. Алгоритмической основой работы Musite является метод опорных векторов (от англ. Support Vector Machine, или SVM). Три параметра были выбраны авторами для репрезентации последовательностей: 1) KNN (K Nearest Neighbor) scores; 2) Disorder scores; 3) Частоты нахождения аминокислот в области известных сайтов фосфорилирования. Параметры 1 и 2 были также использованы авторами iPhos-PseEn, и их краткое описание дано ниже. Выборка, используемая для тренировки модели, была скомпилирована на основе различных баз данных, таких как Swiss-Port, P.ELM, PPA и др., однако не была опубликована авторами. Данный аспект значительно осложняет независимую оценку Musite, особенно учитывая, что были взяты последовательности основных модельных эукариотических организмов.

PHOSFER (PHOsfrorylation Site FindER) был разработан с целью предсказания сайтов фосфорилирования в первую очередь растительных белков. Для выбора способа репрезентации последовательности, авторы воспользовались методом, предложенным в [11] методом, заключающимся в кластеризации (fuzzy clustering) аминокислотных дескрипторов, взятых из базы данных AAIndex, и последующим выборе свойст, наименее коррелирующих друг с другом. Таким образом, 544 коэффициента AAIndex были редуцированы до всего лишь 8 как наиболее репрезентативных. В основу тренировочной и тестировочной выборок были взяты фосфопротеомные данные многих модельных эукариотических организмов, однако авторами были введены специальные индексы, регулирующие вклад каждой последовательности в значение функции ошибки в

зависимости от таксономической принадлежности белка (более точно, коэффициенты отражали сходство между последовательностями различных таксонов, вычисленное с использованием BLOSUM62). Другими словами, чем менее сходны белки организма с растительными белками, тем меньший вклад они вносили в функцию ошибки в проессе обучения модели. Это и позволяет считать PHOSFER моделью, ориентированной на работу с белками растений. Так же, как в случае с Musite, восстановить тренировочную выборку не представляется возможным.

Наиболее новой моделью среди данного списка является PhosPredRF. В качестве дескрипторов аминокислотной последовательности были выделены четыре группы: 1) Коэффициенты, основанные на информационной теории - энтропия Шэннона, относительная энтропия Шэннона и разница между данными коэффициентами; 2) Перекрывающиеся свойства - 10 категориальных переменных, таких как заряд, гидрофобность и т.п., закодированные в бинарный 10-мерный вектор; 3) Разнонаправленные 20-мерные векторы, представляющими каждую аминокислоту в форме унитарного двоичного кода (также известного как one-hot encoding); 4) 21-мерные векторы содержащие различные физические характеристики аминокислот. Авторами было проведено сравнение посредством 10-кратной кросс-валидации с более ранними альтернативами, такими как NetPhos [12], PPRED [13], PhosphoSVM [14], в результате которого было показано превосходство данной модели. Выборки для тестирования и тренировки были скомпилированы на основе P.ELM (9.0) и PPA (3.0), однако были взяты не все сайты фосфорилирования, и критерии их отбора не являются прозрачными. Довольно оригинален и выбор самих баз данных, так как P.ELM (9.0) содержит в основном фосфопротеомные данные человека, а PPA (3.0) - A. thaliana. Детали, касающиеся выбранных для сравнения предикторов даны в Таблице 1.

Стоит отдельно отметить некоторые из предикторов, не вошедших в группу моделей, выбранных для сравнения. iPhos-PseEn [15], созданный относительно недавно, основан, как и PhosPredRF, на алгоритме случайных лесов (от англ. random forest; далее RF [16]). RF является мощным алгоритмом, часто применяемым для различных задач классификации и регрессии в вычислительной биологии для, был использован авторами как ensemble-метод (ensemble от англ. ансамбль). В качестве дескрипторов элементов последовательности белка были выбраны: Disorder Score (disorder от англ. неупорядоченность), являющийся характеристикой локальной стабильности белка; K Nearest Neighbor Score (KNNS), основанный на применении метода k-ближайших соседей и описывающий аминокислотное окружение (5 окружающих аминокислотных остатков) каждого элемента последовательности; 20-мерный вектор, содержащий частоты каждого остатка в пределах последовательности; Position Weight Amino Acid Composition (PWAAC), являющейся локальной характеристикой аминокислотного окружения, успешно примененной в [17] для идентификации сайтов фосфорилирования в последовательностях вирусных белков. Авторы провели сравнение предложенного метода на незави-

симой тестовой выборке. Данное сравнение, в частности, показало, что предложенный авторами метод, является более точным, нежели Musite. Однако отсутствие отдельного приложения и возможности работы веб-сервера с более чем пятью последовательностям не позволяют провести независимую оценку работы данной модели.

Еще одним методом, основанным на алгоритме случайных лесов, разработанным параллельно с iPhos-PseEn, и превзошедшим Musite и PHOSFER на независимой тестовой выборке, является RF-Phos [18]. Сервер, указанный авторами в публикации, недоступен, поэтому, как и в случае с iPhos-PseEn, провести независимую оценку модели не предоставляется возможным. Авторами была проведена большая работа по разработке эффективной репрезентации последовательности для использования с алгоритмом RF. Все особенности, среди которых средняя кумулятивная гидрофобность (от англ. Average Cumulative Hydrophobicity), энтропия Шэннона, доступная сольвенту площадь (SASA - Solvent Accessible Surface Area) и другие, были сконкатенированы в вектора с размерностью в 593. Таким образом, каждая последовательность были представлена массивной $L \times 593$ матрицей, где L - длина последовательности. Исследовательская ценность работы состоит также в том, что авторами были проанализированы использовавшиеся для репрезентации особенности и выделены те, которые в наибольшей степени способствовали точной классификации (в частности, все вышеназванные входят в данную выборку).

Авторы PPRED [13] подошли к проблеме репрезентации иным способом. Вместо дескрипторов, отражающих те или иные физические свойства аминокислот, была использована эволюционная информация, отражающая консервативность тех или иных аминокислотных остатков. Её можно отразить, проведя множественное выравнивание последовательности интереса против выбранной протеомной базы данных. Таким образом, для каждой последовательности с использованием PSI-BLAST были получены PSSM-профили (Позиционная весовая матрица, от англ. Position-Specific Scoring Matrix), широко применяемые в биоинформатике для анализа нуклеотидных и белковых последовательностей, и отражающие относительные частоты замен элементов последовательности. При анализе полученных данных, было сделано наблюдение, что в окружении сайтов фосфорилирования PSSM-профили последовательностей имеют сходства. В качестве алогоритма для обучения модели был использован метод опорных векторов. Модель была разработана для киназо-неспецифичных предсказаний, и доступна через веб-сервер, который, однако способен к однократной обработке лишь одной последовательности.

С технической точки зрения все представленные выше модели полагаются на выбранные авторами вручную способами представления последовательности белка. Другими словами, разработчику модели необходимо принять решение в пользу того или иного варианта репрезентации, и данное решение не может быть названо тривиальным. Ко всему прочему, требуется проведение значительного объема работы по пре-

добработке данных и анализу полученных результатов (с целью выявления наиболее эффективного варианта репрезентации).

В то же время, конволюционные нейронные сети (CNN - Convolutional Neural Networks) способны к автоматическому позиционно-инвариантному извлечению локальных особенностей (в данном случае, в области сайта фосфорилирования) в тренировочных данных. Уже была показана их эффективность при решении разного рода задач, возникающих как в вычислительной биологии, так и за её пределами. Так, с использованием конволюционных нейронных сетей были созданы модели по предсказанию контактной карты (от англ. contact map) белка [19], его вторичной структуры [20, 21] и ферментативной функции [22]. Помимо этого, CNN с большим успехом применяются в области компьютерного зрения, анализа изображений, и даже для моделирования биологического зрения [23].

Хотя, как таковые, сайты модификации являются локальными особенностями последовательности белка, хорошая модель, с целью уменьшения количества ложно - положительных результатов, должна учитывать и более общие особенности первичной структуры, то есть охватывать максимально широкую область аминокислотной последовательности. Таким образом, задача определения сайтов фосфорилирования на самом деле является задачей последовательной классификации, сродни определению вторичной структуры белка или же классификации частей речи предложения в области обработки естесственного языка (от англ. NLP - Natural Language Processing). Тем не менее, ни один из вышеописанных предикторов не содержит в своей основе алгоритмов, ориентированных на последовательную классификацию элементов. Таковыми, в свою очередь, являются рекурентные нейронные сети (от англ. RNN - Recurrent Neural Networks). Более того, как было показано в [24] на примере определения субклеточной локализации белка, гибридные CNN-RNN модели справляются с поставленной задачей лучше, чем каждая из архитектур по отдельности.

3 Материалы и методы

3.1 Используемые данные

В работе были использованы следующие базы данных: PhosphoSitePlus (PSP) [25], Phospho.ELM (P.ELM) [26], dbPPT [27], dbPSP [28] и dbPAF [29]; их краткое описание дано в Таблице 2.

Таблица 2: Количественное описание баз данных (кроме dbPSP).

	PSP	P.ELM9.0	dbPAF	dbPPT
Кол-во сайтов	360728	57390	483001	82175
Кол-во последовательностей	56126	11175	54148	31012
Кол-во ссылок на PMID	-*	2569	15154	78
Среднее кол-во ссылок в расчете на сайт	6.54	1.26	3.74	2.02

*Данные отсутствуют в публичном доступе.

Данные для dbPAF, dbPSP и dbPPT, как утверждается авторами в [27–29], были извлечены главным образом из публикаций в PubMed посредством поиска по ключевым словам, и затем проанализированы экспертами в области. dbPPT содержит около 80 тысяч сайтов фосфорилирования в 30 тысячах белках растений 20 различных видов. dbPSP содержит исключительно прокариотические данные, в то время как в dbPAF агрегированы сайты фосфорилирования основных модельных организмов, таких как H. sapiens, D. melonogaster, S. cerevisae (отдельно стоит отметить отсутствие последовательностей растительных белков). В Phospho.ELM содержатся данные о сайтах фосфорилирования в белках преимущественно позвоночных животных, большинство из которых принадлежат человеку. Три из вышеуказанных ресурсов (кроме dbPSP) предоставляют лишь ссылки на публикации в виде PMID, на основе которых был сделан вывод о наличии сайта фосфорилирования, не указывая, какого рода исследования лежали в основе его обнаружения. PSP же, напротив, не предоставляет PMID, однако содержит информацию об экспериментальных методах в виде трех классов: LT, MS и MS CST (Рис. 3).

3.2 Предобработка данных

На основании пяти данных ресурсов были созданы три различных выборки. В силу выбранной архитектуры модели (а именно, наличия конволюционных слоев; см. ниже), для длин последовательностей были заданы нижняя и верхняя границы. Аминокис-

Рис. 2: Количество сайтов фосфорилирования*, находящееся в базе данных PhosphoSitePlus на момент написания работы, агрегированное в соотвествии с экспериментальными методами.

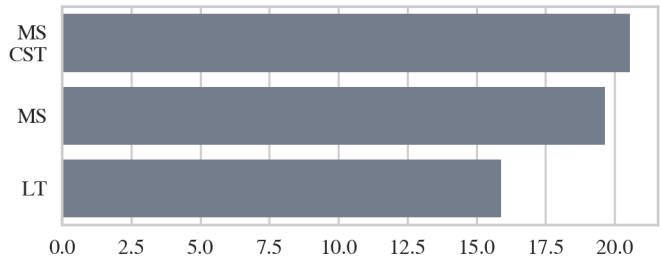


Рис. 3: MS CST - Mass Spectrometry Cell Signaling Technology [30]. MS - Mass Spectrometry. LT - Low Throughput, сборная категория для низкопропускных методов. *Суммарное количество завышено, в силу того, что один и тот же сайт может быть определен разными методами.

лотные последовательности каждой из выборок были прокластеризованы при помощи cd-hit [31], и из каждого кластера, руководствуясь максимальным количеством сайтов фосфорилирования, было отобрано по одной последовательности. Детальное описание выборок находится в Таблице 3.

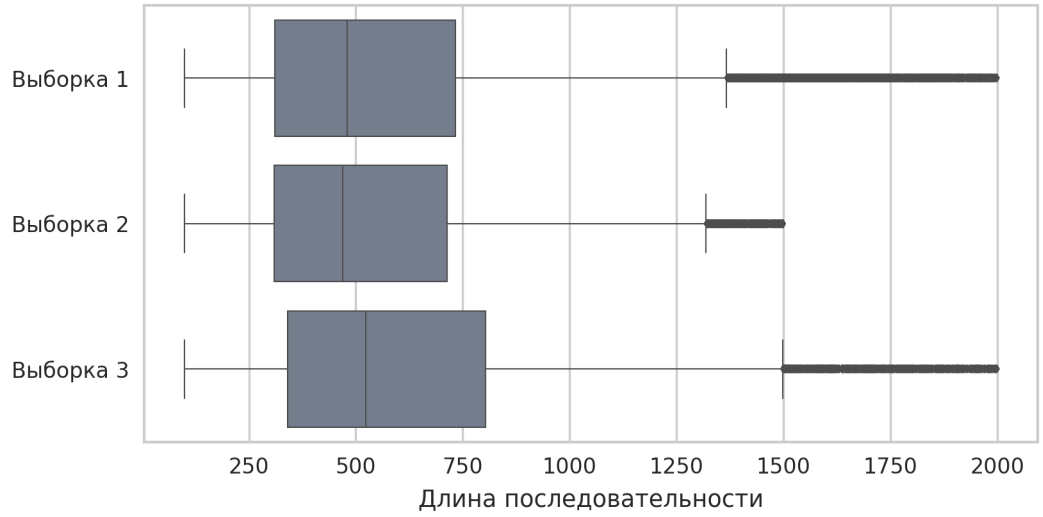
Таблица 3: Детали формирования выборок.

Выборка	I	II	III
Ограничение длины последовательности	100-2000	100-1500	100-2000
Кол-во последовательностей*, 10^3	85/66	42/30	39/25
Процент идентичности для кластеризации	90	95	70
Базы данных	dbPAF; dbPPT; dbPSP	PSP	PSP; P.ELM9.0; dbPAF; dbPPT

*До/После кластеризации

Из выборки II, основанной лишь на PSP, были исключены все сайты фосфорилирования, имевшие менее одной ссылки на низкопропускные методы и менее трех ссылок на высокопропускные. Данный выбор был сделан на основании рекомендаций, опубликованных на ресурсе `PhosphoSitePlus.org`. Аналогичным образом, из выборки III были исключены сайты, имеющие менее трех ссылок в виде PMID (кроме данных из dbPSP, где PMID отсутствуют: в этом случае применение подобных критериев было невозможно, поэтому были взяты все прокариотические последовательности). От каждой из выборок в случайном порядке было отобрано 5% последовательностей с целью созда-

Рис. 4: Диаграмма типа box plot, иллюстрирующая распределение длин последовательностей в каждой из выборок.



ния тестовых наборов данных для сравнения эффективности разрабатываемой модели. Так как GPS3.0 и PhosPredRF были натренированы на подмножестве сайтов фосфорилирования, взятых их P.ELM, все соответствующие последовательности из тестовых выборок были исключены.

Все аминокислоты в последовательностях белков были закодированы целыми числами от 1 до 21. Несмотря на способность рекурентных нейронных сетей работать с последовательностями произвольной длины, из соображений вычислительной эффективности, а также в целях включения в архитектуру конволюционных слоев, все недостающие до максимального значения длины элементы были представлены нулевыми значениями. Таким образом, все последовательности в каждой из сформированных выборок были объединены в единую матрицу с целыми числовыми значениями в диапазоне от 0 до 21. Так как распределение длин последовательностей заметно скошено (Рисунок 4), с целью повышения вычислительной эффективности, и уменьшения разреженности данных, внесенной добавлением нулевых значений, последовательность была разделена с использованием скользящего окна размером N с шагом s. При оценкие работы модели на тестовой выборке все результаты объединялись, после чего вычислялось арифметическое среднее полученных моделью вероятностей в перекрывающихся областях.

Процесс предобработки данных был выполнен с использованием языка `Python3.5` преимущественно с использованием библотек `Pandas` и `Numpy`. В работе по исследованию используемых данных были также использованы технология `Jupyter Notebook` и язык программирования `R`.

3.3 Архитектура модели

Характер задачи определения сайтов фосфорилирования относится к той группе задач, которая в зарубежной литературе известна как seq2seq (sequence-to-sequence): для каждого элемента последовательности на входе нейронная сеть вычисляет вероятность его принадлежности к положительному классу на выходе. Другими словами, входом и выходом для модели служат последовательности числовых значений.

Каждый из элементов последовательности, подаваемой модели на вход, кодируется в вектор определенной длины: так называемый "эмбеддинг" (от англ. embedding - вложение). То есть, каждому аминокислотному остатку, как уникальному классу, соответствует представление в K-мерном векторном пространстве (где K является гиперпараметром), причем конкретные значения каждого из векторов являются изменяемыми в процессе обучения модели параметрами. Таким образом, вместо того, чтобы выбирать некоторые особенности аминокислот, кажущиеся репрезентативными, вручную, модель в автоматической манере подстраивает их внутреннее представление таким образом, чтобы последнее максимально эффективно соответствовало решаемой задаче.

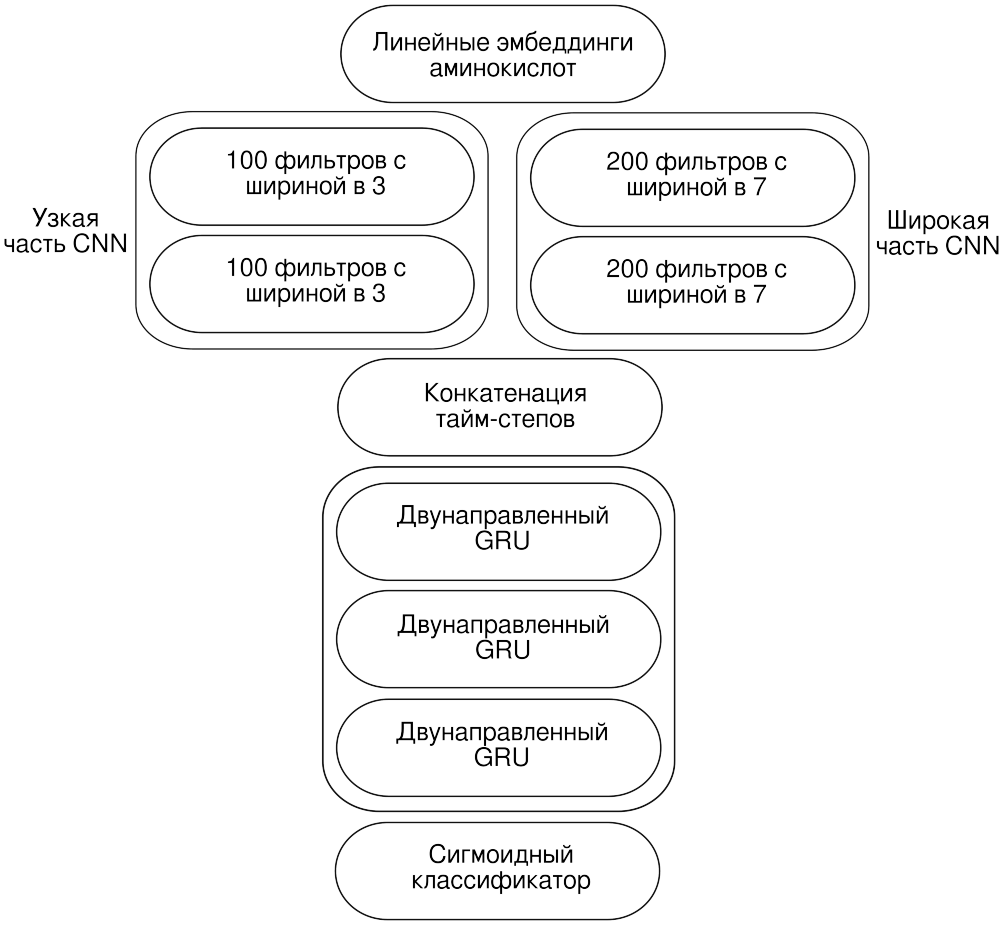
Векторные эмбеддинги затем в параллельной манере подавались в две группы одномерных конволюционных слоев (1D CNN), каждая из которых характеризовалась различным размером и количеством конволюционных фильтров. Размер в одной из групп превосходил размер в другой, что позволяло проводить извлечение особенностей последовательности в двух различных масштабах.

Для обработки извлеченных особенностей при помощи конволюционных слоев были добавлены рекуррентные слои (RNN). Последние, в отличие от стандартных слоев нейронной сети, представляют собой последовательную модель, тренируемую посредством обратного распространения ошибки (back-propagation) через серию временных шагов (time-steps) [19]. Не столь давно были разработаны расширения рекуррентных нейронных сетей, способные к учету долговременных зависимостей, а в частности, LSTM (Long Short-Term Memory) [32] и GRU (Gated Recurrent Unit) [33]. После экспериментирования с обеими вышеназванными архитектурами (LSTM и GRU), выбор был сделан в пользу GRU как не уступающей LSTM по части эффективности, но содержащей меньшее количество параметров.

Для решения проблемы переобучения, были применен метод регуляризации, известный как Dropout [34]. Инициализация весов была проведена по методу Glorot Uniform [35] с использованием рекомендованных параметров. В качестве функции потери была использована бинарная кросс-энтропия, $H(p,q) = -\sum_x p(x) \log q(x)$ а её оптимизация происходила с использованием алгоритма Adam [36].

Код модели был выполнен с использованием Python3.5 и библиотек Keras 2 [37] и TensorFlow 1.5 [38]. Тренировка моделей происходила на видеокартах Titan X и GTX1080 с использованием CUDA.

Рис. 5: Архитектура модели.



На первом этапе в эмбеддинг-слое аминокислотная последовательность обретает внутреннее представление в виде набора векторов с изменяемыми в процессе тренировки параметрами. Затем в параллельной манере происходит подача эмбеддингов в CNN-слои с фильтрами размером 7 и 3. После извлечения локальных особенностей последовательности, происходит конкатенация результатов обеих частей CNN, а результат конкатенации затем обрабатывается рекуррентной частью. Последняя представлена тремя двунаправленными GRU-слоями, последовательно обрабатывающими серию тайм-степов в прямом и обратном направлениях. За рекуррентной частью модели следует выходной слой в виде бинарного классификатора с сигмоидной функцией активации.

3.4 Метод оценки работы модели

В силу значительного дисбаланса положительных и отрицательных классов в сторону последних в качестве основной метрики для оценки работы модели был выбран F_1-score.

$$TPR = \frac{TP}{TP + FN}$$

TPR (от англ. True Positive Rate) вычисляется как отношение числа верно предсказанных положительных классов (TP от англ. True Positive) к общему числу числу поло-

жительных классов - $TP+FN$ (FN от англ. False Negative).

$$PRC = \frac{TP}{TP+FP}$$

PRC (от англ. Precision) вычисляется как отношение числа верно предсказанных положительных классов (TP) к общему числу классов, предсказанных, как положительные ($TP+FP$, где FP от англ. False Positive).

$$F_1 = \frac{2*TPR*PRC}{TPR+PRC}$$

F_1-score является частным случаем F-метрики в случае бинарной классификации, и вычисляется как гармоническое среднее между точностью (PRC) и чувствительностью (sensitivity, или TPR).

4 Результаты

4.1 Выбор гипер-параметров

В процессе работы с общей архитектрурой модели, описанной выше, было протестировано несколько наборов гипер-параметров, По причине большого количества возможных комбинаций, а также учитывая время, необходимое для тренировки модели, проведенные эксперименты нельзя назвать исчерпывающими.

Как бы то ни было, значительный вклад в эффективность работы модели внесло включение в модель конволюционных слоев, следующих за входным слоем. Сравнение с эквивалентной по количеству слоев моделью, где последние представлены только рекурентными слоями, продемонстрировало значительное повышение скорости вычислений без какого-либо ущерба точности предсказаний. Положительно сказалось и разделение конволюционных слоев на две независимые группы: с узкими (покрывающими 3 элемента последовательности) и широкими (покрывающими 7 элементов последовательности) фильтрами. Как было указано выше, в качестве регуляризации был использован метод Dropout. Его суть заключается в выключении нейронов с определенной вероятностью, которая, в свою очередь, является гипер-параметром. Для последней следующие значения были найдены эффективными: 0.3 для конволюционных слоев и 0.1 для рекуррентных.

В эмбеддинг слое, следующим за входным слоем и осуществляющим реализацию внутреннего представления последовательности, гипер-параметром является размерность пространства, в которое помещается последовательность. Как было указано в выше, некоторые из современных моделей имеют крайне массивные представления, достигающие размерности более 500. В результате экспериментирования с размерностью представления, эффективные его значения соответствовали 20.

Были испробованы различные параметры размера N и шага s скользящего окна для каждого из наборов данных. Так, эффективные значения N оказались равными 120, 160, 100 для выборок I-III, соответственно. Эффективные значения s варьировали в пределах 5-10.

В силу того, что количество сайтов фосфорилирования для Ser, Thr и Tyr неодинаково, их численный вклад при вычислении функции ошибки неравнозначен. Значительно на результат работы модели повлияло введение численных коэффициентов - весов, вычисляемых динамически для каждой группы подаваемых во входной слой последовательностей. Для более сбалансированного назначения весов было применено их логарифмическое сглаживание в соотвествии с функцией $f(weight, base) = \log_{base}(weight + base - 1)$. Всем аминокислотным остаткам, кроме Ser, Thr и Tyr, при вычислении функции ошибки были присвоены нулевые значения.

4.2 Эффективность модели

В силу того, что восстановить тренировочную выборку для GPS3.0, Musite и PHOSFER на основе соответствующих публикаций было невозможно, результаты данных предикторов на тестовых выборках, скорее всего, являются завышенными. С другой стороны, Musite был натренирован на выборке, в которой отсутствовали последовательности прокариотических белков как таковые, а PHOSFER задумывался авторами как предиктор, ориентированный на работу с последовательностями белков растений. Также отдельно стоит отметить, что Musite является киназо-специфичным предиктором, и натренирован на соответствующей выборке. Причина, по которой Musite был использован для сравнения, заключается в возможности выбора модели “Eukaryote.ser.thr.tyr”, которую авторы рекомендуют использовать для киназо-неспецифичных предсказаний.

Результаты сравнения работы предикторов на тестовых выборках представлены в Таблице 4.

4.3 Доступность модели

Так как целью данной работы являлось создание доступного пользователю приложения, весь программный код был выложен в открытом репозитории на ресурсе GitHub [39]. Приложение может быть использовано посредством коммандной строки, и было протестировано на операционной системе Ubuntu Linux 16.10. Возможен выбор одной из трех моделей (по умолчанию используется модель, натренированная на выборке II, как показавшая науилучшие результаты), для каждой из которых возможены три режима работы: тренировка, оценка и, собственно, предсказание. Таким образом, пользователь может как адаптировать для своих целей путем её перетренировки на собственной выборке, так и, при наличии необходимых знаний, модифицировать архитектуру и изменять гипер-параметры модели.

Таблица 4: Точность (*PRC*), чувствительность (*TPR*) и F_1-score, соответственно, вычисленные на тестовых выборках наборов данных I-III (описание дано в Материалах и методах). Модель, созданная в работе, обозначена как CNN-RNN.

	CNN-RNN	GPS 3.0*	PhosPredRF	MUSITE*	PHOSFER*
			Dataset I		
Bacteria	0.14/0.13/0.13	0.02/0.99/0.03	0.02/0.09/0.03	0.06/0.10/0.08	0.02/0.37/0.03
Eukaryotes	0.31/0.33/0.32	0.04/0.99/0.07	0.05/0.16/0.09	0.19/0.26/0.22	0.06/0.70/0.11
Metazoans	0.28/0.33/0.30	0.04/0.99/0.08	0.07/0.14/0.09	0.21/0.26/0.23	0.07/0.61/0.13
Mammals	0.28/0.31/0.29	0.04/0.99/0.09	0.08/0.13/0.10	0.22/0.23/0.23	0.08/0.58/0.13
Human	0.29/0.22/0.25	0.05/0.99/0.10	0.08/0.12/0.10	0.22/0.17/0.19	0.08/0.51/0.14
			Dataset II		
Eukaryotes	0.33/0.39/0.36	0.03/0.99/0.07	0.06/0.64/0.11	0.19/0.24/0.21	0.06/0.57/0.11
Metazoans	0.35/0.47/0.40	0.03/0.99/0.07	0.06/0.64/0.11	0.19/0.24/0.21	0.06/0.57/0.11
Mammals	0.39/0.48/0.43	0.04/0.99/0.07	0.06/0.64/0.11	0.19/0.24/0.21	0.06/0.57/0.11
Human	0.36/0.50/0.42	0.05/0.99/0.10	0.08/0.57/0.14	0.24/0.17/0.20	0.08/0.50/0.14
			Dataset III		
Eukaryotes	0.26/0.42/0.32	0.03/0.99/0.06	0.05/0.78/0.09	0.17/0.32/0.23	0.05/0.73/0.10
Metazoans	0.26/0.42/0.32	0.03/0.99/0.05	0.05/0.77/0.09	0.17/0.36/0.23	0.05/0.70/0.09
Mammals	0.26/0.42/0.32	0.03/0.99/0.06	0.05/0.73/0.09	0.18/0.32/0.23	0.05/0.65/0.10
Human	0.39/0.32/0.36	0.03/0.99/0.05	0.05/0.70/0.09	0.18/0.30/0.22	0.05/0.62/0.10

* Результаты могут быть завышены (подробности в тексте).

Для повышения эффективности работы модели рекомендуется использование графических процессоров с 8GB VRAM, поддерживающих CUDA, а также наличие порядка 32GB оперативной памяти для проведения вычислений на больших массивах данных.

5 Обсуждение

Несмотря на возможное преувеличение результатов некоторых предикторов, сравнение работы на всех тестовых выборках (см. Таблицу 4) демонстрирует значительное превосходство созданной модели по значению F_1-score.

Результаты работы всех моделей, выбранных для сравнения, кроме Musite, далеки от наилучших. Налицо отсутствие баланса между значениями *PRC* и *TPR*. Так, значение *TPR* в случае GPS3.0 достигает 99%, однако значение *PRC* не превышает 10%, что объясняет такие низкие показатели F_1-score. Тем удивительнее результаты предска-

заний для белков человека, показанные PHOSFER, более высокие, нежели таковые у предикторов филогенетически более общего назначения GPS3.0 и PhosPredRF.

Наиболее низкие результаты, что справедливо для всех сравниваемых моделей, были показаны для прокариотических последовательностей. Возможно, это связано с отсутствием возможности фильтрации данных по количеству ссылок на экспериментальные источники, а также с совокупностью отличия последовательностей белков про- и эукариот и меньшим количественным их вкладом в выборку I (около 4.5%). Все модели имеют наилучшие показатели при работе с последовательностями белков человека. Созданная модель, а также Musite, являются исключением: максимум их значений F_1-score был достигнут в рамках более общей филогенетической группы млекопитающих. Однако даже в этом случае F_1-score созданной модели более чем в 2 раза превосходит таковой для Musite. В целом, практически все модели имеют наилучшие результаты на выборке II. Именно при формировании данной выборки были применены наиболее строгие критерии отбора сайтов фосфорилирования, что указывает на необходимость их применения при подготовке данных. Это также имплицирует важность наличия подробных сопроводительных аннотаций для записей в базах данных, посвященных ПТМ, что расширило бы возможности по созданию качественных выборок для тренировки моделей, подобных выполненной в данной работе.

6 Выводы

Задача точного определения сайтов фосфорилирования вычислительными методами не является простой сама по себе и при этом значительно осложняется отсутствием возможности введения строгих критериев на этапе формирования выборки. Базы данных продолжают интенсивно расти, поэтому введение подробных сопроводительных аннотаций в значительной мере повысило бы эффективность создаваемых вычислительных моделей, ориентированных на предсказания сайтов пост-трансляционных модификаций, что было успешно продемонстрировано в данной работе.

Существующие вычислительные методы предсказания сайтов фосфорилирования на сегодняшний момент не отличаются высокой точностью работы, а также по ряду причин, подчеркнутых в данной работе, являются устаревшими. Несмотря на многочисленные успешные применения методов глубокого обучения в различных областях вычислительной биологии, поиски актуальных на сегодняшний момент предикторов, основанных на данных методах, не увенчались успехом. По этой причине была предложена гибридная модель, основанная на новейших результатах в области глубокого обучения - конволюционных и рекуррентных нейронных сетях. Созданную модель отличает ряд важных особенностей, среди которых неспецифичность по отношению к определенной группе киназ или филогенетической группе, высокая вычислительная эффективность, простота в использовании, открытость и модифицируемость. Её пре-

восходство было продемонстрировано путем сравнения её с существующими решениями на независимых тестовых выборках, сформированных на основе наиболее свежих экспериментальных данных. Это же, в свою очередь, показывает высокую эффективность примененных методов, что позволяет предположить плодотворное их применение для создания моделей, ориентированных на предсказания и других сайтов ПТМ. Отдельно стоит отметить способность нейронных сетей обучаться без выделения каких-либо репрезентативных особенностей в тренировочных выборках. Созданная модель была дополнена интерфейсом в виде коммандной строки, а весь сопутствующий программный код и созданные выборки данных были опубликованы в открытом репозитории на ресурсе GitHub [39].

Список литературы

[1] Tejaswita M. Karve and Amrita K. Cheema. Small Changes Huge Impact: The Role of Protein Posttranslational Modifications in Cellular Homeostasis and Disease. Journal of Amino Acids, 2011:1–13, 2011.

[2] Fatima Ardito, Michele Giuliani, Donatella Perrone, Giuseppe Troiano, and Lorenzo Lo Muzio. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). International Journal of Molecular Medicine, pages 271–280, 2017.

[3] Philip Cohen. The role of protein phosphorylation in human health and disease: Delivered on June 30th 2001 at the FEBS meeting in Lisbon. European Journal of Biochemistry, 268(19):5001–5010, 2001.

[4] Susanne B. Breitkopf Asara and John M. Determining in vivo Phosphorylation Sites using Mass Spectrometry. NIH Public Access, pages 1–27, 2013.

[5] Omictools.org. `https://omictools.com/phosphorylation-sites-category`. Accessed: 2017-10-17.

[6] Martina Audagnotto and Matteo Dal Peraro. Protein post-translational modifications: In silico prediction tools and molecular modeling. Computational and Structural Biotechnology Journal, 15:307–319, 2017.

[7] Yu Xue, Jian Ren, Xinjiao Gao, Changjiang Jin, Longping Wen, and Xuebiao Yao. GPS 2.0, a Tool to Predict Kinase-specific Phosphorylation Sites in Hierarchy. Molecular & Cellular Proteomics, 7(9):1598–1608, 2008.

[8] Jianjiong Gao, Jay J. Thelen, A. Keith Dunker, and Dong Xu. Musite, a Tool for Global Prediction of General and Kinase-specific Phosphorylation Sites. Molecular & Cellular Proteomics, 9(12):2586–2600, 2010.

[9] Brett Trost and Anthony Kusalik. Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. Bioinformatics, 29(6):686–694, 2013.

[10] Leyi Wei, Pengwei Xing, Jijun Tang, and Quan Zou. PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. IEEE Transactions on NanoBioscience, 1241(c):1–1, 2017.

[11] Indrajit Saha, Ujjwal Maulik, Sanghamitra Bandyopadhyay, and Dariusz Plewczynski. Fuzzy clustering of physicochemical and biochemical properties of amino Acids. Amino Acids, 43(2):583–594, 2012.

[12] Nikolaj Blom, Steen Gammeltoft, and Søren Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of Molecular Biology, 294(5):1351–1362, 1999.

[13] A.K. Biswas, N. Noman, and A.R. Sikder. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. BMC Bioinformatics, 11, 2010.

[14] Yongchao Dou, Bo Yao, and Chi Zhang. PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. Amino Acids, 46(6):1459–1469, 2014.

[15] Wang-Ren Qiu, Xuan Xiao, Zhao-Chun Xu, Kuo-Chen Chou, Wang-Ren Qiu, Xuan Xiao, Zhao-Chun Xu, and Kuo-Chen Chou. iPhos-PseEn: Identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget, 5(0), 2016.

[16] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. Random forests: From early developments to recent advancements. Systems Science and Control Engineering, 2(1):602–609, 2014.

[17] Shu Yun Huang, Shao Ping Shi, Jian Ding Qiu, and Ming Chu Liu. Using support vector machines to identify protein phosphorylation sites in viruses. Journal of Molecular Graphics and Modelling, 56:84–90, 2015.

[18] Hamid D. Ismail, Ahoi Jones, Jung H. Kim, Robert H. Newman, and Dukka B. Kc. RF-Phos: A Novel General Phosphorylation Site Prediction Tool Based on Random Forest. BioMed Research International, 2016, 2016.

[19] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model, volume 13. 2017.

[20] Chao Fang, Yi Shang, and Dong Xu. MUFold-SS: Protein Secondary Structure Prediction Using Deep Inception-Inside-Inception Networks. pages 2–7, 2017.

[21] Akosua Busia, Jasmine Collins, and Navdeep Jaitly. Protein Secondary Structure Prediction Using Deep Multi-scale Convolutional Neural Networks and Next-Step Conditioning. pages 1–10, 2016.

[22] Evangelia I Zacharaki Corresp, Corresponding Author, and Evangelia I Zacharaki. Prediction of protein function using a deep convolutional neural network ensemble.

[23] Nikolaus Kriegeskorte. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. Annual Review of Vision Science, 1(1):417–446, 2015.

[24] Søren Kaae Sønderby, Casper Kaae Sønderby, Henrik Nielsen, and Ole Winther. Convolutional LSTM networks for subcellular localization of proteins. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9199:68–80, 2015.

[25] Peter V. Hornbeck, Jon M. Kornhauser, Sasha Tkachev, Bin Zhang, Elzbieta Skrzypek, Beth Murray, Vaughan Latham, and Michael Sullivan. PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Research, 40(D1):261–270, 2012.

[26] Holger Dinkel, Claudia Chica, Allegra Via, Cathryn M. Gould, Lars J. Jensen, Toby J. Gibson, and Francesca Diella. Phospho.ELM: A database of phosphorylation sites-update 2011. Nucleic Acids Research, 39(SUPPL. 1):261–267, 2011.

[27] H. Cheng, W. Deng, Y. Wang, J. Ren, Z. Liu, and Y. Xue. dbPPT: a comprehensive database of protein phosphorylation in plants. Database, 2014(0):bau121–bau121, 2014.

[28] Zhicheng Pan, Bangshan Wang, Ying Zhang, Yongbo Wang, Shahid Ullah, Ren Jian, Zexian Liu, and Yu Xue. dbPSP: a curated database for protein phosphorylation sites in prokaryotes. Database : the journal of biological databases and curation, 2015(September 2017):bav031, 2015.

[29] Shahid Ullah, Shaofeng Lin, Yang Xu, Wankun Deng, Lili Ma, Ying Zhang, Zexian Liu, and Yu Xue. dbPAF: an integrative database of protein phosphorylation in animals and fungi. Scientific reports, 6(March):23534, 2016.

[30] Cellsignal.com. `https://www.cellsignal.com`.

[31] Weizhong Li and Adam Godzik. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13):1658–1659, 2006.

[32] Sepp Hochreiter and J Urgen Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8):1735–1780, 1997.

[33] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. pages 1–9, 2014.

[34] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15:1929–1958, 2014.

[35] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), 9:249–256, 2010.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. pages 1–15, 2014.

[37] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[38] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[39] Github.com. `https://github.com/skoblov-lab/KUPPNet`.